一种考虑风力作用的 KNN 城市 AQI 预测算法 *

杨丰玉，王宝英†，陈　英，冯　涛，陈涛苹

(南昌航空大学 软件学院，南昌 330063)

摘　要：考虑风力对城市空气质量指数(AQI)的重要影响，基于 KNN 算法提出一种新的模型对城市 AQI 进行预测。该模型主要依赖于数据间的局部相似性和依赖性，再将风力因素对城市 AQI 的影响进行量化并加入到 KNN 预测结果中，得到最终预测结果。实验对九个重点城市进行 AQI 预测，结果表明，该模型相较传统 KNN 方法预测得到的 AQI 值，准确率大幅度提升，对城市 AQI 的预测具有指导意义。

关键词：空气质量指数；K 近邻；风力因素；预测

中图分类号：TP183　　doi: 10.3969/j.issn.1001-3695.2017.12.0797

K nearest neighbor urban forecasting algorithm considering wind factors

Yang Fengyu, Wang Baoying†, Chen Ying, Feng Tao, Chen Taoping

(*School of Software*, *Nanchang Hangkong University*, *Nanchang* 330063, *China*)

Abstract: Considering the important influence of wind on air quality index(AQI) , this paper proposes a new model , which based on KNN algorithm to predict the urban AQI. The model mainly relies on the local similarity and dependence between data, and quantifies the impact of wind factors on urban AQI and adds it to the KNN forecast results to get the final forecast results. The experimental results of AQI in nine major cities show that the proposed model has significantly improved the accuracy of AQI compared with the traditional KNN method. The model has guiding significance for the prediction of urban AQI .

Key words: Air quality index; K nearest neighbor; Wind factors; forecast

0　引言

空气质量情况与人们的工作生活有极大的关系，空气中的污染物不利于人体健康，易导致呼吸道疾病。空气污染严重时，肉眼可见度极低，易发生道路交通事故及影响航班出行，严重影响人们的日常生活和社会经济发展。

2016 年 12 月，法律法规网(http://www.lc123.net/xw/rd/2016-12-08/607831.html)指出,我国 78.4%的城市空气质量不达标，部分地区冬季重度及以上污染[1]频发。有效的城市空气质量指数预测至关重要，人们可以根据空气质量指数预测来安排出行计划及方式，相关部门也可根据空气质量指数预测在污染严重时提醒人们。故高准确率的空气质量指数预测很有必要。

1　时间序列预测常用方法

时间序列预测方法主要分为基于模型的时间序列预测和基于人工智能的时间序列预测两类。基于模型的时间序列预测方法主要有 ARMA 模型和 ARIMA 模型[2]。基于人工智能的时间序列预测方法主要有 SVM 和神经网络。2014 年，席超等人[3]使用 ARMA 模型对导航卫星钟差进行预测，与多项式模型和灰色模型相比，ARMA 模型预测导航卫星钟差准确率更高。2017 年，Zhang 等人[4]使用 ARMA 和粒子群优化的核极限学习机（KELM）构建预测模型，对电价进行预测，澳大利亚和西班牙的市场实验结果表明该方法能有效预测电价。2016 年，翟静等人[5]使用 ARIMA 和 BP 神经网络对中国粮食产量进行预测，结果表明预测结果较准确。2017 年，Ruby-Figueroaa 等人[6]使用 ARIMA 模型对超滤果汁六小时的渗透通量进行预测，分别对使用不同渗透膜的佛手柑、猕猴桃和石榴汁的渗透通量进行实验，结果表明预测效果理想。2015 年，刘爱国等人[7]使用优化核函数的 SVM 对超短期风电功率进行预测，该方法取得较好的预测结果。2017 年，Direito 等人[8]使用多通道高维特种集和多类支持向量机对癫痫患者的发病进行预测，该方法在预测过程中定义了一组区分患者发病前和发病后的特征，实验结果表明该方法对患者发病预测有效。2017 年，李博等人[9]使用景气预测的先行指标和神经网络的适应性等特点，构建四种预测

收稿日期：2017-12-07；**修回日期：**2018-01-25　　**基金项目：**江西省自然科学基金资助项目（20161BAB212034）

作者简介：杨丰玉（1980-），男，江西九江人，副教授，硕士，主要研究方向为软件工程、大数据；王宝英（1992-），女（通信作者），江西上饶人，硕士，主要研究方向为大数据（403247067@qq.com）；陈英（1982-），男，副教授，博士，主要研究方向为软件工程、模式匹配；冯涛（1993-），男，江西赣州人，硕士，主要研究方向为大数据；陈涛苹（1993-），男，江西鹰潭人，硕士，主要研究方向为大数据.

模型对北京市中关村高新科技园总收入进行预测，实验结果表明四种给予神经网络的预测模型均适用于景气预测。2017 年，Asencio-Cortés 等人[10]使用从文献中得到的地震指标，基于神经网络构建预测模型，对东京 2015 年高于 5 级、时间跨度大于 7 天的地震进行预测，实验结果表明该方法有效。

KNN 算法是模式识别的一种，适用于非线性序列数据。算法核心是数据间的依赖性：如果训练数据与查询数据相近（该训练数据称为查询数据的相近邻），则认为相近邻的预测数据与该查询数据的预测数据相近，即若查询数据相近，则预测数据也相近，相近程度越高则相似度越大。为确保预测数据不因某个特殊点而造成极大误差，选用 K 个最近邻来获得最终的预测结果。到目前为止，KNN 算法已成功应用于多个领域。

2014 年， Zheng 等人[11]用 KNN 算法进行短期车流量预测，实验结果表明该算法适用于各地复杂的车流量预测。2017 年，Oliveira 等人[12]用 KNN 算法预测短期用水需求，在训练数据较多的情况下，该算法取得了较好的结果。AQI 序列数据与车流量及用水需求量一样均为非线性序列数据，且经观察空气质量有春冬采暖季污染严重、夏天一般空气质量较好的规律，数据间具有局部相似性和依赖性。故而选择 KNN 算法应用于 AQI 预测。

KNN 算法总共分为以下四个步骤。

a)选取 n 组查询数据 X_i 。

$$X_i=\{x_i,x_{i+1},x_{i+2},\cdots,x_{i+n-2},x_{i+n-1}\} \tag{1}$$

b)在训练数据中依次选取 n 组训练数据 X_j

$$X_j=\{x_j,x_{j+1},x_{j+2},\cdots,x_{j+n-2},x_{j+n-1}\} \tag{2}$$

c)计算 X_i 和 X_j 之间的欧氏距离 dis 。

$$dis=\sqrt{\left(X_j-X_i\right)^2} \tag{3}$$

d)选择欧式距离最小的 k 组数据，加入权重，计算得到预测值。

2 本文提出的方法

2.1 风力因素的影响

风对空气质量有很大的影响，风能将污染物吹散或带离某个地区，能带走多少污染物或带来多少污染物取决于风力的大小。风向所影响的是风将污染物带向哪个方向或从哪个方向带来。考虑污染一般在城市出现，城市与城市之间的地区产生污染的可能性很小，故风对城市空气质量产生的影响一般是将污染物带离城市，将城市周边质量较好的空气带到城市；且风向对空气质量的影响还需结合各城市周边山脉等环境。影响小且情况复杂，故本方法不考虑风向对空气质量的影响，只考虑风力对空气质量的影响。且考虑到 AQI 变化的原因有多种，而风对 AQI 造成的影响一般是下降，若上升则考虑为其他因素，本方法不考虑其他因素，故只考虑 AQI 下降情况，将后一天 AQI 值与当前 AQI 值的差值作为当前风力对 AQI 的影响，忽略当日风对当日 AQI 值的影响。

2.2 算法步骤

1)选取 K 近邻

①初始化两个长度为 k 的数组，其中一个数组存放 k 个最小的欧氏距离，数组为 $dis[k]$，且该数组按升序排序。数组中最后保留的最小欧式距离一般在 100 左右，故先将数组中的每个值初始化为 9999，对最终结果不会产生任何影响，$dis[k]=\{9999,9999,9999,\cdots,9999,9999,9999\}$；另一个数组存放对应于欧式距离最小的训练数据的日期即，数组为 $flag[k]$ 。

②选预测数据前 n 天数据作为的查询数据 X_i ，$X_i=\{x_i,x_{i+1},x_{i+2},\cdots,x_{i+n-2},x_{i+n-1}\}$ 。

③在该指定日期 AQI 预测实验过程中，若是第一次获取训练数据，则执行第④步；否则，执行第⑤步。

④从训练数据第一条数据起依次获取连续 n 天 AQI 数据作为 X_j ， $X_j=\{x_j,x_{j+1},x_{j+2},\cdots,x_{j+n-2},x_{j+n-1}\}$ 。

⑤从训练数据指定数据起依次获取连续 n 天 AQI 数据作为 X_j ， $X_j=\{x_j,x_{j+1},x_{j+2},\cdots,x_{j+n-2},x_{j+n-1}\}$ 。

⑥计算 X_i 和 X_j 之间的欧氏距离 dis ， $dis=\sqrt{\left(X_j-X_i\right)^2}$ 。

⑦分别比较 dis 与 $dis[k]$ 中每个数的大小，若 $dis\le dis[i]$，即 X_i 和 X_j 之间的欧式距离小于或等于 $dis[k]$ 中的某个数，则将 $dis[k]$ 中欧氏距离最大的数据删除并将 dis 插入到相应位置；与此同时，更新 $dis[k]$ 数组相应位置的值。

⑧据欧氏距离最小的 k 个值 $dis[k]$ 和训练数据对应日期 $flag[k]$，取得相应的 k 个预测 AQI 值 $FAQI[k]$ 作为预测值的 K 近邻。

2)加入权重

⑨训练数据日期往后移一天，若仍可依次获取 n 条 AQI 数据，则跳转至第④步；否则，本次指定日期的 AQI 预测筛选过程结束，跳转至第⑩步。

⑩在 K 近邻中加入权重(欧氏距离越小，权重越大)。k 组欧氏距离之和为 $Dis=\sum dis$，每组欧氏距离最小的值相应的预测值分别为 $FAQI[i]$ 。则加权 AQI 预测值计算公式如下：

$$AQI=\sum FAQI[i]*\frac{Dis-FAQI[i]}{(k-1)*Dis} \tag{4}$$

⑪若测试数据均预测完毕，则实验步骤结束；否则预测日期往后移一天，并跳转至第①步。

3)考虑风力因素的 AQI 预测值

⑫计算下一天 AQI 值 AQI_{t+1} 与当前 AQI 值 AQI_t 的差值 D_t 作为当天风力对下一天 AQI 的影响。

$$D_t=AQI_{t+1}-AQI_t \tag{5}$$

⑬筛选指定风力值 x 下差值为负即 $D_t<0$ 的值 $D_{x.net}$，计算差值对应当天实际 AQI 的百分比 $d_{x.t}$,即当前风力对当天 AQI 所造成的影响。

$$d_{x.t}=\frac{D_{x,net}}{AQI_t}\times 100\% \tag{6}$$

⑭计算该风力下 m 组 $d_{x.t}$ 对 AQI 所造成影响的平均值 d_x。

$$d_x=\frac{1}{m}\sum d_{x.t} \tag{7}$$

⑮使用计算所得的风力影响值对 KNN 预测所得结果 AQI_t 进行修正。根据测试数据上一天的风力值 x 引入风力影响值，得到修正后的 AQI 值 AQI_{t1}。

$$AQI_{t1}=AQI_t\times(1+d_x) \tag{8}$$

2.3 实验

1)实验数据来源及划分

实验中所用到的实验数据均从网站“天气后报”(http://www.tianqihoubao.com/aqi/)取得。由于本算法是一种具有普遍性的算法，故共取全国东西南北中各地九个重点城市的天气数据作为研究对象，包括日期、AQI、风力等数据；九个城市分别为：北京、西安、合肥、南昌、南京、上海、武汉、长沙、郑州。各城市所获取数据均从 2013-10-28 至 2017-12-31，共 1526 条数据，分为训练数据和测试数据两个部分。其中，2013-10-28 至 2016-12-31 的 1161 条数据作为训练数据，2017-1-1 至 2017-12-31 的 365 条数据作为测试数据。

2)实验数据预处理

在“天气后报”中获取的数据中，有一部分数据的 AQI 信息缺失，故而需要对其进行预处理。且在风力数据中，风力信息种类繁多，也需要对风力数据进行预处理。

①缺失数据

将所有缺失数据的 AQI 均设为 9999，以保证包含该天数据的欧氏距离计算所得到的数值足够大，故而包含该天的数据不会被选中作为查询数据的 K 近邻。

②风力数据

选取进行研究的九个城市对象中，风力因素最小的均为 3 级，故而 3 级风不予考虑对 AQI 的影响，从 3 级以上风开始考虑。不考虑风向因素，只考虑风力等级因素，将风力因素数字等级化；数字化方式：①若风力等级为 3，则直接数字化为 3；若风力等级为 3-4，则直接数字化为 3.5；以此类推。②在风力数据数字化过程中，若包含两种等级的风力，则直接将风力等级高的数字化。

3)算法参数设置

本算法需要确定两个参数，分别是查询数据个数 n 和最近邻的个数 k。在九个实验城市中随机选取三个城市(选取城市为：南昌、武汉和西安)。①根据不同的查询天数 n($n\in[3,10]$)分别对三个城市 2017 年 8 月 AQI 进行预测，实验结果如表 1 所示。查询天数 $n=5$ 时，三个城市的平均误差均最小，故取 $n=5$。②根据相关研究，K 近邻的个数 $k\in[30,60]$ 时，预测效果最好[13]，但该理论建立在数据量达 10000 的基础上。由于实验数据有限，只有 1000 条，k 值过大，反而会影响预测效果，故本算法取 $k\in[3,10]$。根据不同的 K 近邻个数($k\in[3,10]$)分别对三个城市 2017 年 8 月 AQI 进行预测，实验结果如表 2 所示。在 $k=7$ 时效果较好，故取 $k=7$。

表 1　不同查询天数 n 所得预测结果的平均绝对值误差

城市	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10
南昌	0.251	0.238	0.245	0.254	0.258	0.257	0.267	0.264
武汉	0.387	0.354	0.304	0.353	0.339	0.358	0.371	0.340
西安	0.342	0.293	0.300	0.326	0.306	0.318	0.342	0.313

表 2　不同 K 近邻个数所得预测结果的平均绝对值误差

城市	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10
南昌	0.266	0.276	0.251	0.255	0.244	0.245	0.241	0.234
武汉	0.312	0.309	0.304	0.327	0.330	0.328	0.325	0.319
西安	0.296	0.305	0.300	0.303	0.307	0.306	0.298	0.296

4)实验结果评估方法

对只使用 KNN 方法预测得到的 AQI 值 AQI_t 以及考虑风力作用后得到的 AQI 值 AQI_{t1} 进行评估。 分别计算 AQI_t 和 AQI_{t1} 与当天实际 AQI 值 AQI_r 的绝对误差 E_t 和 E_{t1} 以及 E_t 和 E_{t1} 的平均值 MAE_t 和 MAE_{t1}：

$$E_t=\frac{|AQI_t-AQI_r|}{AQI_r}\times 100\% \tag{9}$$

$$E_{t1}=\frac{|AQI_{t1}-AQI_r|}{AQI_r}\times 100\% \tag{10}$$

$$MAE_t=\sum E_t \tag{11}$$

$$MAE_{t1}=\sum E_{t1} \tag{12}$$

3　实验结果及分析

3.1　实验结果

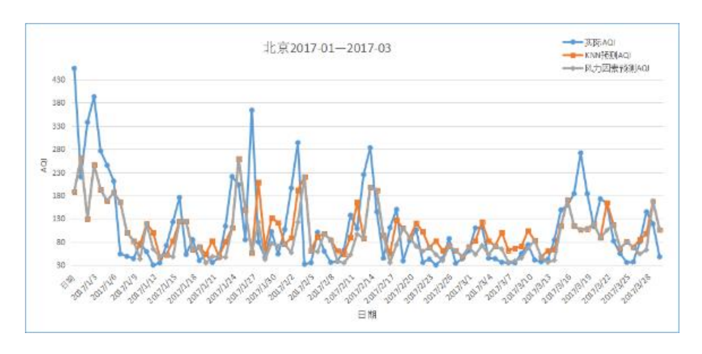
1) 三个典型城市实际 AQI 和两次预测结果对比

本文实验首先选取中、西、北三个方位各一个典型城市作为实验对象。图 1～3 是三个典型城市 2017 年实际 AQI 值、预测 AQI 值以及引入风力后预测 AQI 值之间的对比效果。每张图显示三个月的数据；图 4 是三个典型城市 2017 年 8 月有风天气实际 AIQ、KNN 预测 AQI 和本文算法预测 AQI 的对比簇状柱形图。由图 1～4 可知，虽然有部分有风天气引入风力因素后预测 AQI 不如 KNN 预测 AQI 准确，但影响较小，整体效果上看，引入风力因素对 AQI 预测具有积极作用。

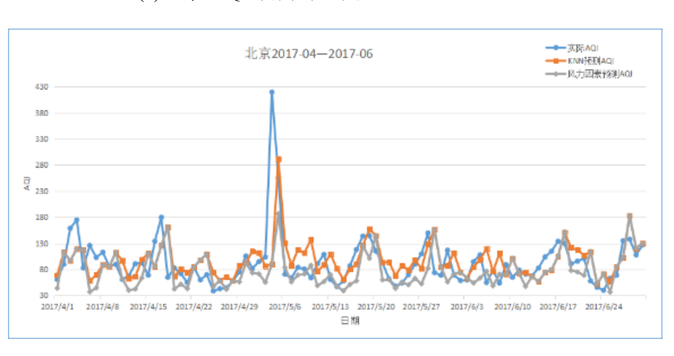
2) 风力因素对 AQI 预测的影响

如表 3 所示，是风力因素对九个城市 2017 年 AQI 预测的准确率的影响。如图 5(a)所示，是九个城市引入风力因素前后对有风天气 AQI 预测的误差对比簇状柱形图； (b)是九个城市引入风力因素前后对整月 AQI 预测的误差对比簇状柱形图。由表 3 和图 5 可知，引入风力因素后，每个城市的 AQI 预测准确率均得到提升。合肥、南京准确率提升幅度最大，分别达到 18.75% 和 12.50%，提升效果最为明显；预测效果也最为明显。西安、武汉和南昌提升幅度不明显，但这三个城市本身预测效果已经

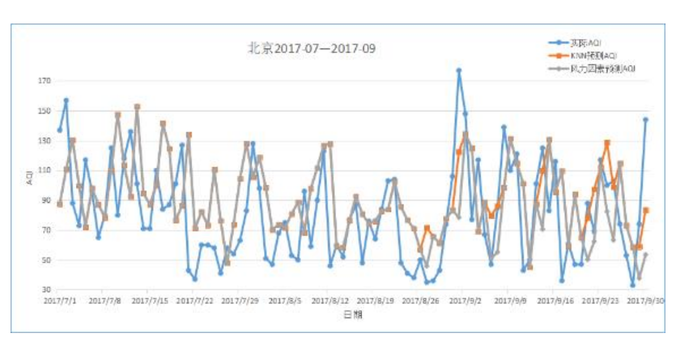
很好，有风天数少，准确率提升幅度虽不明显，但总体预测效果很好。郑州、长沙和上海准确率提升效果较好，这三个城市本身 KNN 预测效果不差，故而引入风力因素后 AQI 预测效果较好。北京 KNN 预测效果较差，预测准确率提升效果一般，仅 5.88%，故预测效果不理想。



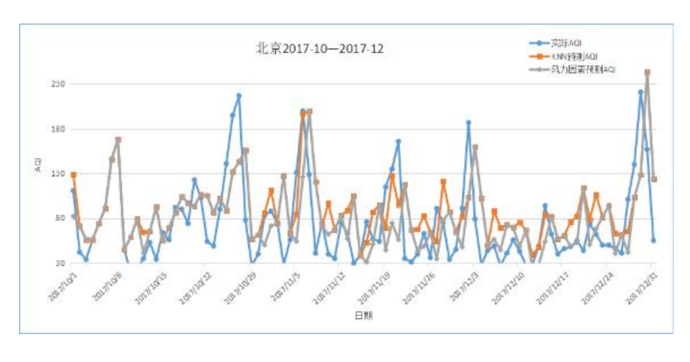
(a)北京 AQI 预测对比图 2017-01—2017-03



(b)北京 AQI 预测对比图 2017-04—2017-06

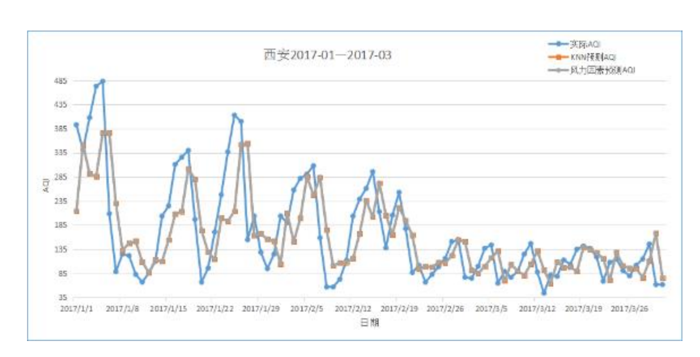


(c)北京 AQI 预测对比图 2017-07—2017-09

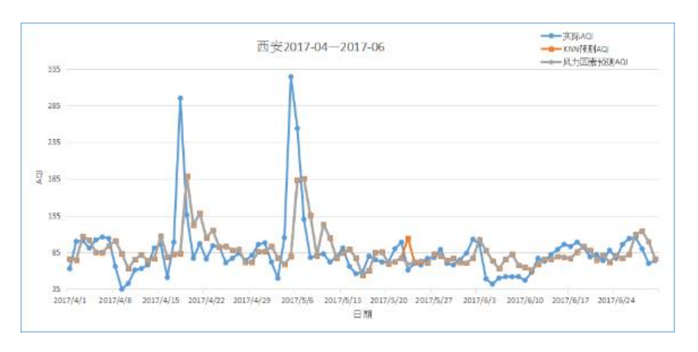


(d)北京 AQI 预测对比图 2017-10—2017-12

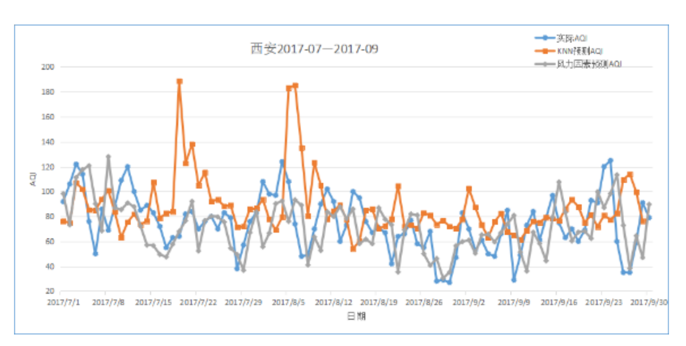
图 1　北京 2017 年实际 AQI、KNN 预测 AQI 和风力因素预测 AQI 折线对比



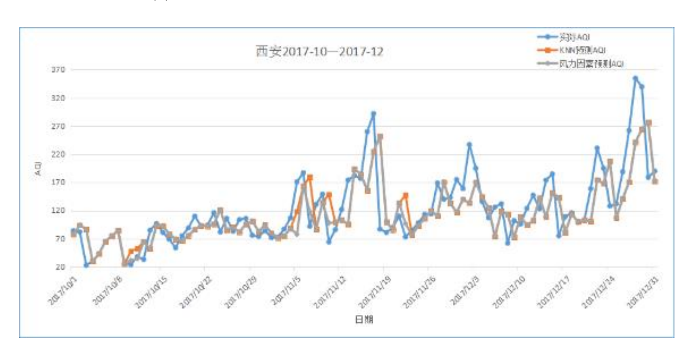
(a)西安 AQI 预测对比图 2017-01—2017-03



(b)西安 AQI 预测对比图 2017-04—2017-06

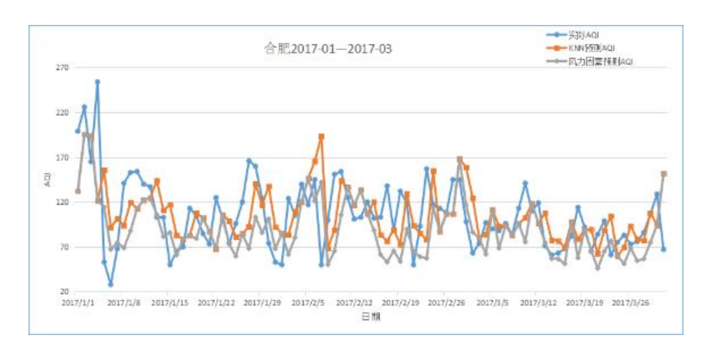


(c)西安 AQI 预测对比图 2017-07—2017-09

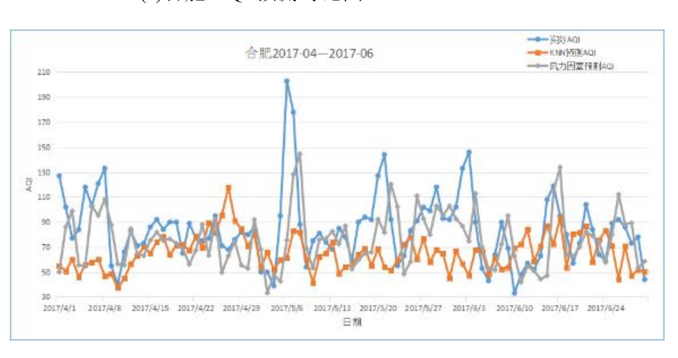


(d)西安 AQI 预测对比图 2017-10—2017-12

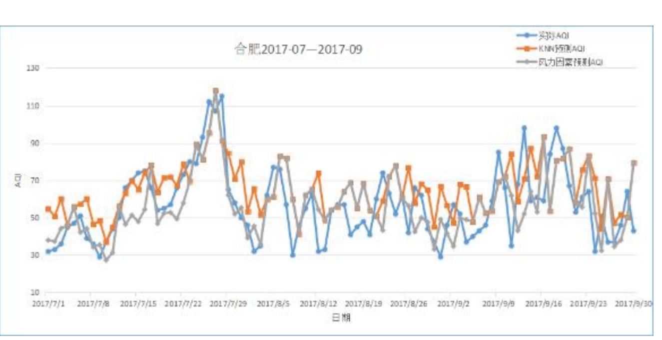
图 2　西安 2017 年实际 AQI、KNN 预测 AQI 和风力因素预测 AQI 折线对比



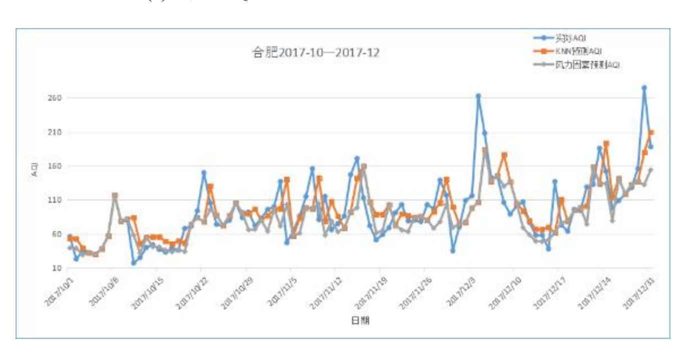
(a)合肥 AQI 预测对比图 2017-01—2017-03



(b)合肥 AQI 预测对比图 2017-04—2017-06

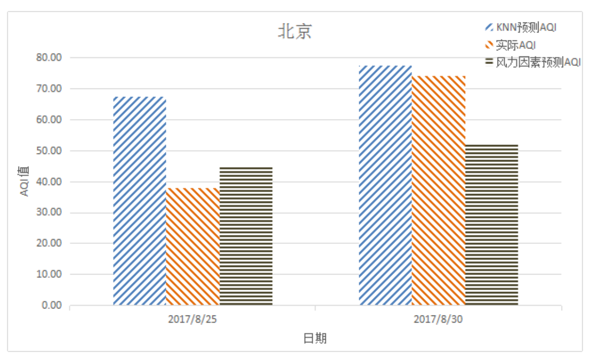


(c)合肥 AQI 预测对比图 2017-07—2017-09

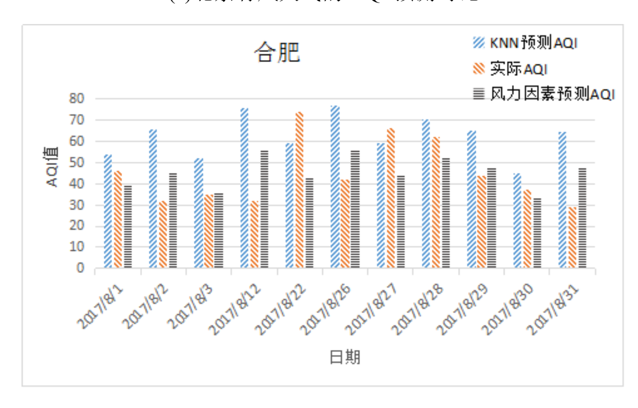


(d)合肥 AQI 预测对比图 2017-10—2017-12

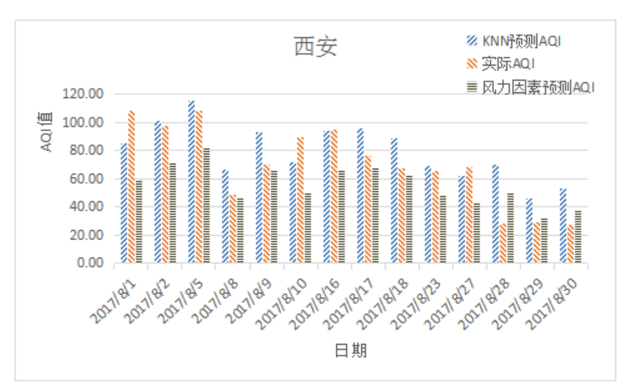
图 3　合肥 2017 年实际 AQI、KNN 预测 AQI 和风力因素预测 AQI 折线对比



(a)北京有风天气的 AQI 预测对比



(b)合肥有风天气的 AQI 预测对比

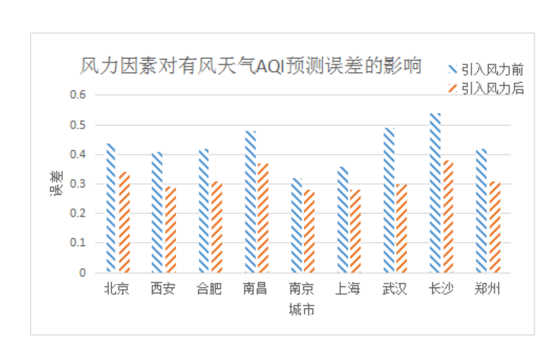


(c)西安有风天气的 AQI 预测对比

图 4　三个典型城市有风天气实际 AQI、KNN 预测 AQI 及风力因素预测 AQI 对比

表 3　风力因素对 AQI 预测准确率的影响

城市	有风天数	提升准确率天数	降低准确率天数	KNN 预测误差	风力因素预测误差	准确率提升百分比
北京	107	58	75	0.51	0.48	5.88%
合肥	192	98	94	0.32	0.26	18.75%
西安	38	21	17	0.30	0.29	3.33%
南昌	62	29	33	0.31	0.29	6.45%
南京	345	148	197	0.32	0.28	12.50%
上海	115	61	54	0.33	0.31	6.06%
武汉	35	21	14	0.30	0.29	3.33%
长沙	79	45	34	0.35	0.32	8.57%
郑州	107	52	55	0.33	0.30	10.00%



(a)风力因素对有风天气 AQI 预测的影响



(b)风力因素对有总体 AQI 预测影响

图 5　风力因素对有风天气和总体 AQI 预测效果的影响

3）误差分布情况

如图 6 所示，是本文算法对九个城市共 279 天 AQI 预测后的误差分布图。在十个城市共 279 条数据中，引入风力后的预测 AQI 误差中，绝对误差在 $[0,0.2)$ 区间的天数最多，占比 44.62%，随着误差增大，天数逐渐减少，$[0.6,1.0)$ 区间，总占比 7%，。但是在 $[1.0,+\infty)$ 区间，占比又高达 3.93%。本算法对城市 AQI 的预测总体效果很好，但也存在部分数据预测很不准确。

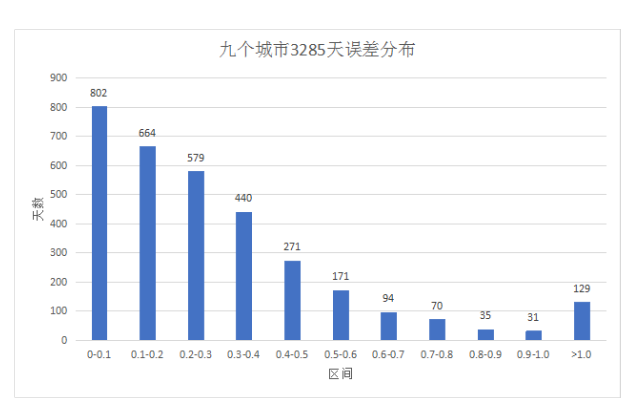


图 6　九个城市 3285 天误差分布

3.2　原因分析

本文提出的方法只考虑数据与数据之间的局部相似性和依赖性及简单的风力因素，实际上 AQI 值受到天气变化、污染物排放以及城市周边环境等多种因素影响。单一考虑数据间的关系以及简单的风力因素，无法完美解决 AQI 复杂的成因问题，故而存在一定误差。

①突变点无法处理

在 AQI 数据中，存在一系列数据，前几天的 AQI 呈下降趋势，且前一天有风。该种情况下，该天 AQI 值较前一天应该下降，但实际 AIQ 值反而上升；或者存在一列数据前几天的 AQI 值呈上升趋势，且前一天无风，该天 AQI 值较前一天应该上升，但实际 AQI 值反而下降，这样的点称之为突变点。突变点一般是因为受到其他因素的影响，可能是天气的变化，也可能污染物排放的变化，但是本文提出的方法只是简单考虑了风力因素，对其他因素产生的影响无法进行量化。

②数据有限

在“天气后报”网站上获取一个城市的历史 AQI 值只有 1 500 条，本文提出的引入风力因素的 KNN 算法主要依赖的是数据间相似性和依赖性，数据量不够大的情况下，筛选得到的最近邻的预测 AQI 值可能会有极大的偶然性误差，从而造成最终的预测结果不够理想，如北京市。

4　结束语

本文在 KNN 算法的基础上，提出引入风力因素的 KNN 算法，对北京、西安、合肥、南昌、南京、上海、武汉、长沙、郑州九个重点城市的 AQI 进行预测，预测结果表明九个城市 KNN 算法预测的 AQI 在引入风力因素后与实际 AQI 的误差均下降。

本文提出的方法主要依赖数据间的相似性和依赖性，但 AQI 变化原因复杂，需要考虑的因素很多，且数据有限，预测 AQI 值的偶然性误差较大，故部分城市预测结果理想，但也有城市预测结果不是很理想。在未来的研究里，将致力于多方位考虑 AQI 变化的原因，使用深度学习或深度学习等方法，对城市 AQI 进行预测，以期获得更好的预测结果。

参考文献：

[1] 杨思琪，赵丽华. 随机森林算法在城市空气质量预测中的应用 [J]. 统计与决策, 2017 (20): 83-86.

[2] 伦淑娴，林健，姚显双. 基于小世界回声状态网的时间序列预测 [J]. 自动化学报, 2015, 42 (9): 1669-1679.

[3] 席超，蔡成林，李思敏，等. 基于 ARMA 模型的导航卫星钟差长期预报 [J]. 天文学报, 2014, 55 (1): 78-88.

[4] Zhang Yang, Li Ce, Li Lian. Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods [J]. Applied Energy, 2017 (190): 291-305.

[5] 翟静，曹俊. 基于时间序列 ARIMA 与 BP 神经网络的组合预测模型 [J]. 统计与决策, 2016 (4): 29-32.

[6] Ruby-Figueroaa R, Saavedrab J, Bahamondec N, *et al.* Permeate flux prediction in the ultrafiltration of fruit juices by ARIMA models [J]. Journal of Membrane Science, 2017 (524): 108-116.

[7] 刘爱国，薛云涛，胡江鹭，等. 基于 GA 优化 SVM 的风电功率的超短期预测 [J]. 电力系统保护与控制, 2015, 43 (2): 90-95.

[8] Direito B, Teixeira C A, Sales F, *et al.* A realistic seizure prediction study

based on multiclass SVM [J]. International Journal of Neural Systems, 2017, 27 (3):

[9] 李博, 王建国, 李静文. 基于神经网络的景气预测模型 [J]. 统计与决策, 2017 (10): 43-46.

[10] Asencio-Cortés G, Martínez-Álvarez F, Troncoso A, *et al.* Medium–large earthquake magnitude prediction in Tokyo with artificial neural networks [J]. Neural Computing and Applications, 2017, 28 (5): 1043-1055.

[11] Zheng Zuduo, Su Dongcai. Short-term traffic volume forecasting: a K-nearest neighborapproach enhanced by constrained linearly sewing principlecomponent algorithm [J]. Transportation Research Part C, 2014, 43: 143-157.

[12] Paulo José A, Dominic L, A M. K-Nearest neighbor for short term water demand forecasting [C]// Proc of World Environmental and Water Resources Congress. 2017: 201-510.

[13] 罗辛, 欧阳元新, 熊璋, 等. 通过相似度支持度优化基于 K 近邻的协同过滤算法 [J]. 计算机学报, 2010, 33 (8): 1440-1441.